

Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances

Arthur B. Kennickell, Federal Reserve Board

Abstract

Donald Rubin has suggested many times that one might multiply impute all the data in a survey as means of avoiding disclosure problems in public-use datasets. Disclosure protection in the Survey of Consumer Finances is a key issue driven by two forces. First, there are legal requirements stemming from the use of tax data in the sample design. Second, there is an ethical responsibility to protect the privacy of respondents, particularly those with small weights and highly salient characteristics. In the past, a large part of the disclosure review of the survey required tedious and detailed examination of the data. After this review, a limited number of sensitive data values were targeted for a type of constrained imputation, and other undisclosed techniques were applied. This paper looks at the results of an experimental multiple imputation of a large fraction of the SCF data using software specifically designed for the survey. In this exercise, a type of range constraint is used to limit the deviations of the imputations from the reported data. The paper will discuss the design of the imputations, and provide a preliminary review of the effects of imputation on subsequent analysis.

Introduction

Typically, in household surveys there is the possibility that information provided in confidence by respondents could be used to identify the respondent. This possibility imposes an ethical, and sometimes a legal, burden on those responsible for publishing the survey: It is necessary to review the data for items that could be highly revealing of the identity of individuals, and to filter the data made available to the public to minimize the degree of disclosure [1]. A recent issue of the *Journal of Official Statistics* (vol. 9, no. 2, 1993) deals with many aspects of this problem.

The Survey of Consumer Finances (SCF) presents two particularly serious disclosure risks. First, the survey is designed to measure the details of families' balance sheets and other aspects of their financial behavior. Second, the SCF oversamples wealthy families. Because of the sensitive nature of the data collected and because the sample contains a disproportionate number of people who might be at well-known, at least in their localities, disclosure review of the SCF is particularly stringent [2].

There is a growing belief that publicly available records, such as credit bureau files, real estate tax data, and similar files make it increasingly likely that an unscrupulous data user might eventually come closer to identifying an SCF respondent [3]. Several protective strategies have been proposed, but many proposals — truncation, simple averaging across cells, random reassignment of data, etc., — raise serious obstacles for many of the analyses for which the SCF is designed. The prospect of either being unable to release any information, or having to alter the data in ways that severely restrict their usefulness makes it imperative that we explore alternative approaches to disclosure limitation.

Most disclosure limitation techniques attempt to release some transformation of the data that preserves what is deemed to be the important information. Taking this idea to one farsighted conclusion, Donald

Rubin has suggested on several occasions creating an entirely synthetic dataset based on the real survey data and multiple imputation (see, e.g., Rubin, 1993) [4]. My impression is that most people have viewed the idea of completely simulated data with at least suspicion [5]. Such an exercise presents considerable technical difficulties. However, even if it is not possible to create an ideal simulated dataset, we may learn something from the attempt to create one. This paper describes several explorations in this direction.

Multiple imputation has played an important role in the creation of the public datasets for the SCF since 1989. In both the 1989 and 1992 surveys, a set of sensitive monetary variables was selected for a set of cases, the responses to those variables were treated as range responses (rather than exact dollar responses) and they were multiply-imputed using the standard FRITZ software developed for the SCF (see Kennickell, 1991). The approach has been broadened in the 1995 survey based on the work reported here. In the experiments discussed in this paper, several approaches are taken to imputing all of the monetary values in the 1995 SCF.

The first section of the paper provides some general information on the content of the SCF and the sample design and gives a review of the past approach to disclosure review. Because of the importance of imputation in the work reported here, the second section reviews the FRITZ imputation model. The third section discusses the special manipulation of the data for this experiment and presents some descriptive results. A final section summarizes the findings of the paper and points toward future work.

The 1995 Survey of Consumer Finances

The SCF is sponsored by the Board of Governors of the Federal Reserve System in cooperation with the Statistics of Income Division of the IRS (SOI). Data collection for the 1995 SCF was conducted between the months of July and December of 1995 by the National Opinion Research Center (NORC) at the University of Chicago. The interviews, which were performed largely in person using computer-assisted personal interviewing (CAPI), required an average of 90 minutes — though some took considerably longer.

Because the major focus of the survey is household finances, the SCF includes questions about all types of financial assets (checking accounts, stocks, mutual funds, cash value life insurance, and other such assets), tangible assets (principal residences, other real estate, businesses, vehicles, and other such assets) and debts (mortgages, credit card debt, debt to and from a personally-owned business, education loans, other consumer loans, and other liabilities). To meet the analytical objectives of the survey, detailed information is collected on every item. For example, for up to six checking accounts, the SCF asks the amount in the account, the owner of the account, and the institution where the account is held. The actual name of the institution is not retained, but a linkage is established to every other place in the interview where the institution is referenced, and detailed questions are asked about the institution. For automobiles, the make, model, and year of the car are requested along with the details of the terms of any loan for the car. Detailed descriptions of types of properties and business that the household owns are collected, along with information on the financial flows to and from the household and the businesses.

To provide adequate contextual variables for analysis, the SCF also obtains data on the current and past jobs of respondents and their spouses or partners, their pension rights from current and past jobs, their marital history, their education, the ages of their parents, and other demographic characteristics. Data are also collected on past inheritances, future inheritances, charitable contributions, attitudes, and many other variables.

Although the combination of such an broad array of variables alone is sufficient cause to warrant intensive efforts to protect the privacy of the individual survey participants, a part of the SCF sample design introduces further potential disclosure problems. The survey is intended to be used for the analysis of financial variables that are widely distributed in the population — e.g., credit card debt and mortgages — and variables that are more narrowly distributed — e.g., personal businesses and corporate stock. To

provide good coverage of both types of variables, the survey employs a dual-frame design (see Kennickell and Woodburn, 1997). In 1995, a standard multi-stage area-probability sample was selected from 100 primary sampling units across the United States (see Tourangeau, et al., 1993). This sample provides good coverage of the variables. A special list sample was designed to oversample wealthy households. Under an agreement between the Federal Reserve and SOI, data from the Individual Tax File (ITF), a sample of individual tax returns specially selected and processed by SOI, are made available for sampling [6].

<i>Data Range</i>	<i>Rounded to Nearest</i>
>1 million	10,000
10,000 to 1 million	1,000
1,000 to 10,000	100
5 to 1,000	10
-5 to -1,000	10
-1,000 to -10,000	100
-10,000 to -1 million	1000
Negative numbers smaller than -1 million truncated at -1 million	
Negative numbers between -1 and -5 unaltered	

The area-probability design raises no particularly troubling issues beyond the need to protect geographic identifiers that is common to most surveys.

However, the list sample raises two distinct problems. First, it increases the proportion of respondents who are wealthy. Such people are likely to be well-known at least in their locality, and because of the relatively small number of such people, it is more likely that data users with malicious intent could match a respondent to external data if sufficient information were released in an unaltered form. Second, because SOI data have been used in the design of the sample, there is a legal requirement that SCF data released to the public be subjected to a disclosure review similar to that required before the release of the public version of the ITF.

Generally, the SCF data have been released to the public in stages. This strategy has allowed us to satisfy some of the most immediate demands of data users, while allowing time to deal with more complex disclosure issues. Once a variable has been released, no amount of disclosure review can retrieve the information, and it can be much more difficult to add variables later because of the possible interactions of sensitive variables. In the past, staged release has allowed users to build a case for including additional variables, and we have been able to accommodate many such requests.

In 1992, the last year for which final data were released at the time this paper was written, the internal data were altered in the following ways for release [7]. First, geography, which was released at the level of the nine Census regions, was altered systematically; observations were sorted and aligned by some key indicators, and geography was swapped across cases.

Second, unusual categories were combined with other related categories — e.g., among owners of miscellaneous vehicles, the categories “boat,” “airplane,” and “helicopter” were combined. Third, a set of cases with unusual wealth or income were chosen, and a random set of other cases was added to the group. For these cases, key variables (for which complete responses were originally given) were multiply imputed subject to range constraints that ensured that the outcomes would be close to the initially reported values. Fourth, a set of other unspecified operations was performed to increase more broadly the perceived uncertainty associated with all variables in every observation; these operations affect both actual data values and the “shadow” variables in the dataset that describe the original state of each variable [8]. As a final step, all continuous variables were rounded as shown in Table 1. Generally, it is impossible to tell with certainty from the variables observed by a user of the public dataset which variables may have been altered and how they were altered.

Table 1.—Rounding of Continuous Variables

A similar strategy is being followed for the 1995 SCF. The one significant change is in the imputation of data for the cases deemed “sensitive” and the random subset of cases described in step three. For the 1995 survey, all monetary data items in the selected cases will be imputed. Depending on the reception of the data by users, this approach may be extended in the 1998 SCF.

FRITZ Imputation Model

Because the principal evidence reported in this paper turns critically on the imputation of monetary variables, it is important to outline some of the more important characteristics of the FRITZ model, which was originally developed for the imputation of the 1989 SCF and has been updated for each round of the survey since then. This discussion focuses on the imputation of continuous variables (see Kennickell, 1991).

Figure 1 shows a hypothetical set of observations with various types of data given. In the figure, “X” represents complete responses, “R” symbolizes responses given as a type of range, and “O” indicates some type of missing value. In the SCF, there is a lengthy catalog of range and missing data responses, and this information is preserved in the shadow variables. Respondents in the 1995 SCF had the option of providing ranges in many ways: as an arbitrary volunteered interval (e.g., between 2,546 and 7,226), as a letter from a range card containing a fixed set of intervals (e.g., range “G” means 5,001 to 7,500), or as the result of answering a series of questions in a decision tree the intervals of which varied by question [9]. Data may be missing because the respondent did not know the answer, refused to answer, because the respondent did not answer a question of a higher order in a sequence, because of recording errors, or other reasons.

The FRITZ system is an iterative multiple imputation model based on ideas of Gibbs sampling. The system acts on a variable-by-variable basis, rather than simultaneously drawing a vector of variables [10]. Within a given iteration, the most generally applied continuous variable routine is, in essence, a type of randomized regression, in which errors are assumed to be normally distributed [11].

One factor that distinguishes the model from the usual description of randomized regression imputation models is the fact that the FRITZ model is tailored to the missing data pattern of each observation. In Figure 1, all of the missing data patterns shown are different, and they are not monotone (Little, 1983). For most continuous variables, the program generates a covariance matrix for a maximal set of variables that are determined to be relevant as possible conditioning variables. For a given case, the model first determines whether a particular variable should be imputed. Given that the variable should be imputed, the FRITZ

model computes a regression for the case using the variables in the maximal set that either are not originally missing or are already imputed within the particular iteration for the case. Finally, the model draws from the estimated conditional distribution until an outcome is found that satisfies any constraints that may apply. Constraints may take several forms. When a respondent has given a range response to a question, FRITZ uses the range to truncate the conditional distribution. Constraints may also involve cross-relationships with other variables, or simply prior knowledge about allowable outcomes. Specification of the constraints is very often the most complex mechanical part of the imputations.

Figure 1. — Hypothetical Missing Data Patterns

<i>Variables</i>										
<i>Observations</i>	X	O	X	X	X	X	X	X	O	X
	O	X	X	X	X	R	X	X	X	X
	X	X	O	O	O	O	X	X	O	R
									
	R	X	X	O	O	X	X	X	X	X
	X	X	X	X	X	X	X	X	R	O
X = reported value R = range value O = missing value										

As noted, once a variable has been imputed, its value is taken in later imputations as if it were originally reported by the respondent. In a given imputation, variables which were originally reported as a range but are not yet imputed within the iteration, are given special treatment. Range reports often contain substantial information on the location of related variables, and one would like to use this knowledge in imputation. In the ideal, it is not difficult to write down a general model that would incorporate many types of location indicators. However, in practice such a model would quickly exhaust the degrees of freedom available in a modestly sized survey like the SCF. In practice, we adopt a compromise solution. Values reported originally as ranges are initialized at their midpoints, and these values are used as conditioning variables for other imputations until the final choice within the range is imputed.

The FRITZ model produces multiple imputations. For simplicity, the strategy adopted is to replicate each observation five times and to impute each of these “implicates” separately. Because different implicates may be imputed to take very different paths through the data, this arrangement allows users to apply standard software to the data.

The iteration process is fairly straightforward. In the first iteration, all the relevant population moments for the imputation model are computed using all available data, including all non-missing pairs of data for the covariance calculations. As imputations progress in that iteration, the covariance estimation is based on increasingly “complete” data. In the second iteration, all population moments are computed using the first iteration dataset, and a new copy of the dataset is progressively “filled in.” In each successive iteration, the process is similar. Generally, the distribution of key imputations changes little after the first few iterations. Because the process is quite time-consuming, the model for the 1995 SCF was stopped after six iterations [12].

Experiments in Imputation for Disclosure Limitation

In this section, I report on three experiments in using multiple imputation for disclosure avoidance (summarized in Figure 2). In these experiments every monetary variable for every observation in the survey was imputed [13]. In the first experiment, all complete reports of dollar values were imputed as if the respondent had originally reported ranges which ran from ten percent above the actual figures to ten percent below that figure. In keeping with our usual practice of using midpoints of ranges as proxies for location indicators in imputation, the original values were retained until the variable was imputed. The second experiment also retained the reported value for conditioning, but imposed no range constraints on the allowed outcomes other than those required for cross-variable consistency. The third experiment treated the original values as if they were completely missing (that is, they were unavailable as conditioning variables) and, like the second experiment, imposed no prior bounds on the imputations; other monetary responses that were originally reported as ranges were also treated as completely missing values for purposes of conditioning, but their imputed values were constrained to lie within the reported ranges.

Figure 2. — Design of Experiments

<i>Experiment</i>	<i>Range Constraints</i>	<i>Use Original Value as Initial Location Indicator</i>
1	±10%	Yes
2	None	Yes
3	None	No

For several reasons, these experiments fall short of Rubin’s ideal that one impute an entire dataset conditioning only on general information — even possibly using only distributional data external to the actual sample. First, the experiments deal only with the dollar variables in the SCF. Second, all complete responses other than monetary responses are used as conditioning variables. Third, the imputations of range responses are constrained to lie within the reported ranges, even in experiment three. Finally — and most probably importantly — the results are specific to the particular specification of the FRITZ model. Inevitably there are deep compromises of theory made in implementing almost any empirical system. For imputation, such compromises may be less pressing when the proportion of missing data is relatively small, as is usually the case in the SCF. These compromises may cause larger distortions when much larger fractions of the data are imputed. A key question in evaluating the results here is how well the system performs under this more extreme condition. Because we also have the originally reported values, it is possible to make a direct evaluation of the performance of the model.

Despite the shortcomings of the three experiments, they seem very much in the spirit of Rubin's proposal. Because the experiments show the effects of progressively loosening the constraints on imputation, I believe the results should provide useful evidence in evaluating the desirability of going further in developing fully simulated data.

The mechanical implementation of these experiments was reasonably straightforward. In the first experiment, the shadow variables of all complete reports of dollar values were set to a value which would normally indicate to the FRITZ model that the respondents had provided distinct dollar ranges. Values equal to the points ten percent above and 10 percent below the reported value were placed in the appropriate positions in a file that the model normally assumes contains such information. In the second and third experiments, a special value was given to the shadow variable to indicate that there were no range constraints on the imputations other than those that enforce cross-variable consistency. In experiments one and two, the initial values of complete responses were left in the dataset at the beginning of imputation; during the course of imputations, these values were used for conditioning until they were replaced by an imputed value, which was used to condition subsequent imputations. In experiment three, values originally reported completely were set to a missing value, and the usual midpoints of range responses were also set to a missing value. Thus, no dollar variables in the third experiment were available for conditioning until they were imputed. In each of the experiments, the imputations were treated as if they were the seventh iteration of the SCF implementation of FRITZ. Thus, estimates of the population moments needed for the model were computed using the final results of the sixth iteration.

In the absence of technical problems — far from the case with the work for this paper for which the imputation system was subject to a massively larger than normal stress — each version of the experiment would require approximately three weeks to run through the entire dataset on a fast dedicated Sun server. More importantly, each execution would also require about 2 gigabytes of disk space for the associated work files. The process could probably be made at least somewhat more efficient, but the time available for debugging such a potentially complex change was limited. A compromise has been adopted here. The first of the eight modules of the SCF application of FRITZ was run for all of the experiments. This module deals largely with total household income and various financial assets.

Figures 3 through 6 show descriptive plots of data from the three experiments for the following four variables: total income, amount in the first savings account, the amount of Treasury bills and other Federal bonds (referred to hereafter as "T-bills"), and the total value of financial assets [14]. The first three of these variables are intended to span a broad set of types distributions; total financial assets, a variable constructed from many components, is included to show the effects of aggregating over the potentially large number of responses to questions about the underlying components. The impression from looking at a broader set of variables is very similar. Each of the figures is divided into two sets of three panels. The top three panels show the distribution for experiments one through three, of the (base-10) logarithm of the originally reported values less the average across the five imputates of the logarithm of the corresponding imputed values ("bias"), where the distribution is estimated as an unweighted average shifted histogram (ASH). The bottom three panels are ASH plots for the three experiments, of the distribution over all cases of the standard deviation of the multiply-imputed values within observations.

For experiment one, the distribution of bias has a mode at approximately zero for all the variables. This is not surprising given that the outcome is based on models estimated using reported data for these observations. In the case of income, savings balances, and T-bills, the distribution of bias is fairly concentrated, with the 10th and 90th percentiles of the distribution corresponding to a bias of only about 5 percent (± 0.02 on the scale shown). The distributions of bias for savings accounts and T-bills are

Figure 3a: ASH Plots of Distribution over Observations of $\text{Log}_{10}(\text{Actual Value}) - \text{Mean}(\text{Log}_{10}(\text{Imputation}))$ within Observations, Total Household Income, Experiments 1-3.

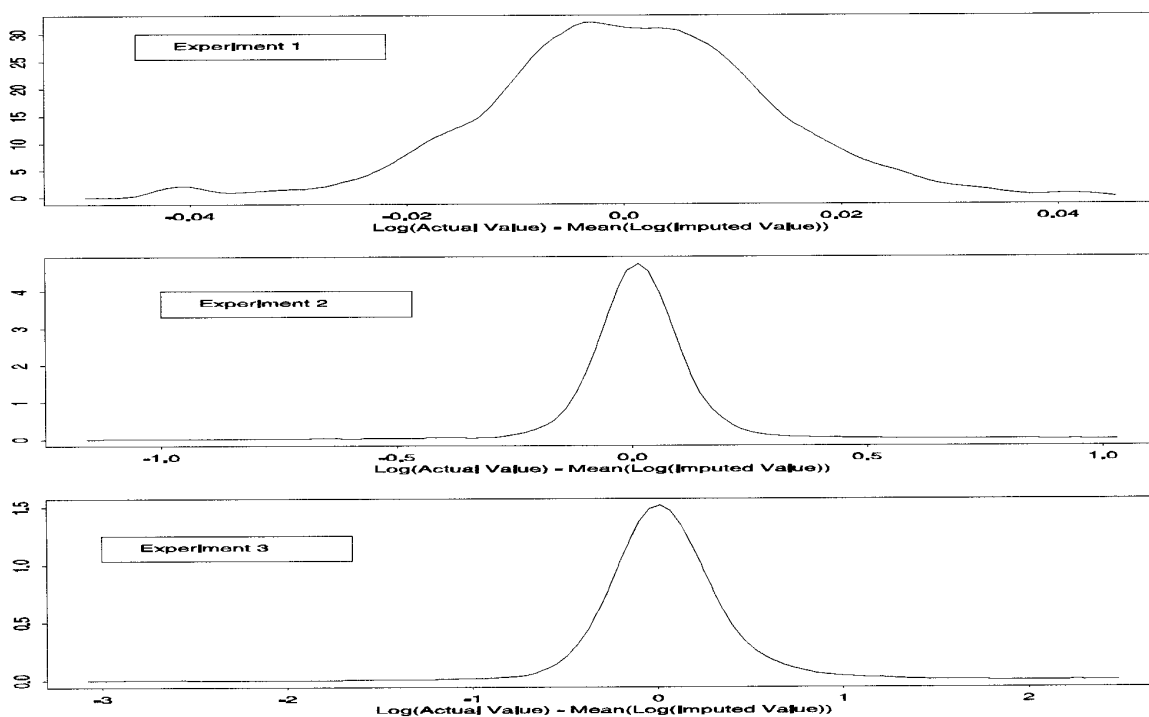


Figure 3b: ASH Plots of Distribution over Observations of the Standard Deviation of $\text{Log}_{10}(\text{Imputation})$ within Observations, Total Household Income, Experiments 1-3.

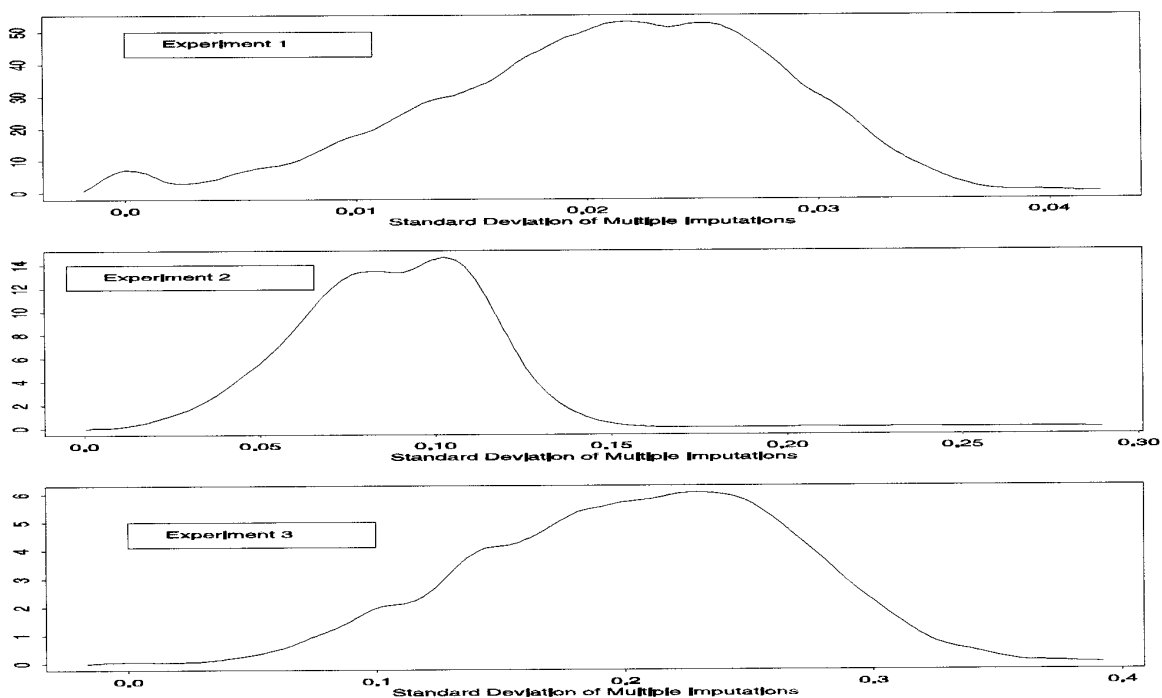


Figure 4a: ASH Plots of Distribution over Observations of $\text{Log}_{10}(\text{Actual Value}) - \text{Mean}(\text{Log}_{10}(\text{Imputation}))$ within Observation, Balance in First Savings Account, Experiments 1-3.

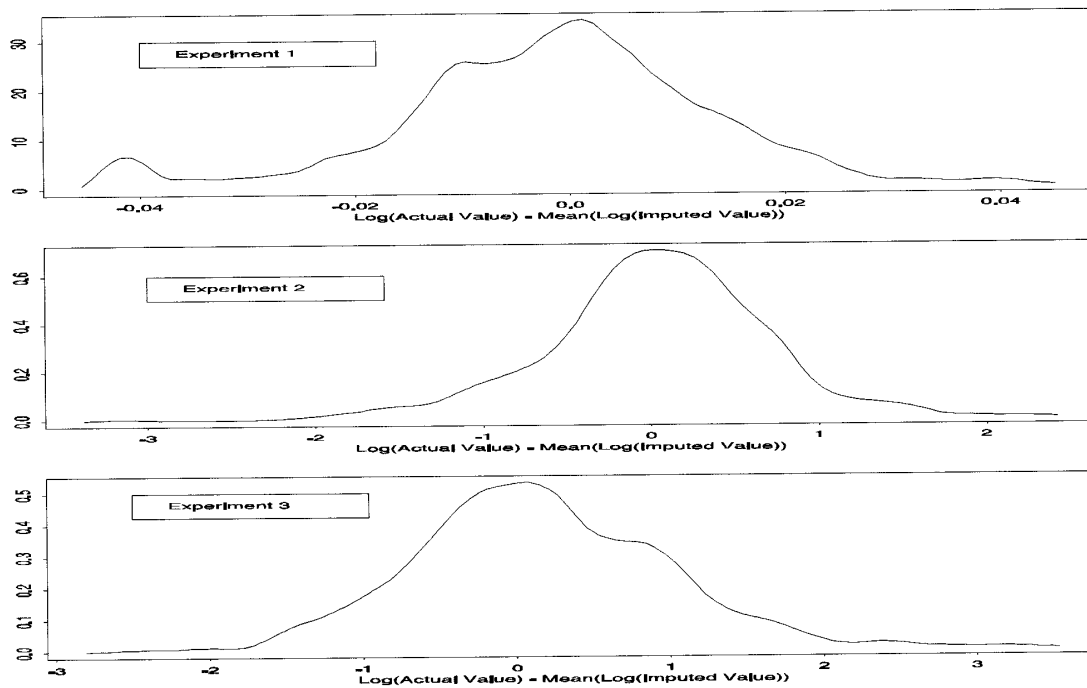


Figure 4b: ASH Plots of Distribution over Observations of the Standard Deviation of $\text{Log}_{10}(\text{Imputation})$ within Observations, Balance in First Savings Account, Experiments 1-3.

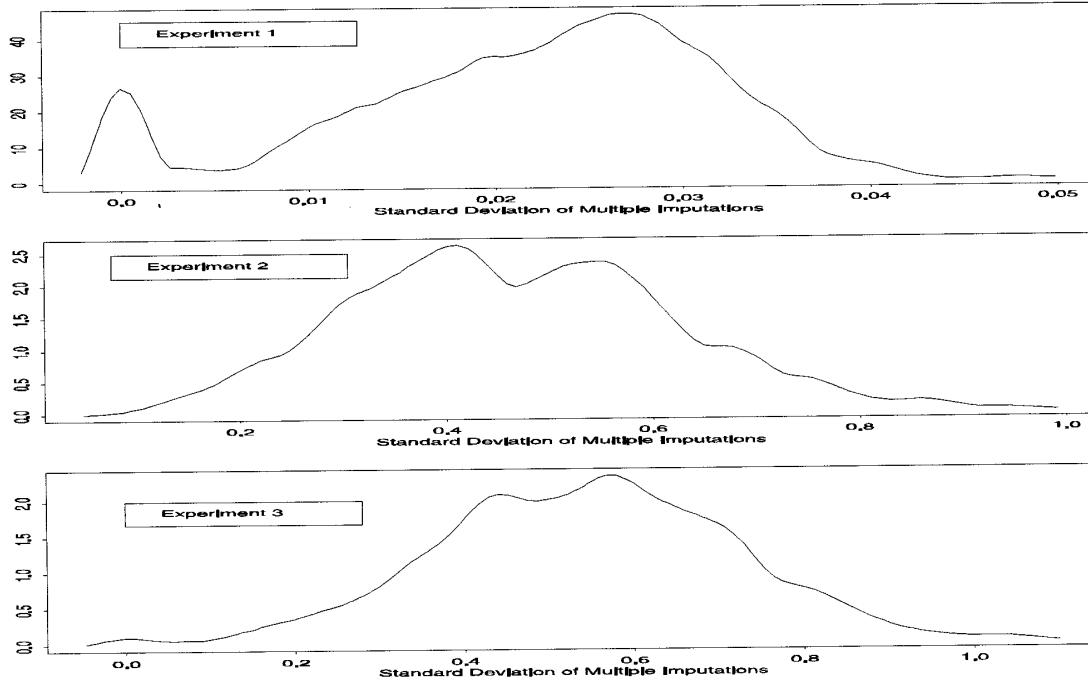


Figure 5a: ASH Plots of Distribution over Observations of $\text{Log}_{10}(\text{Actual Value}) - \text{Mean}(\text{Log}_{10}(\text{Imputation}))$ within Observations, Face Value of T-Bills, Experiments 1-3.

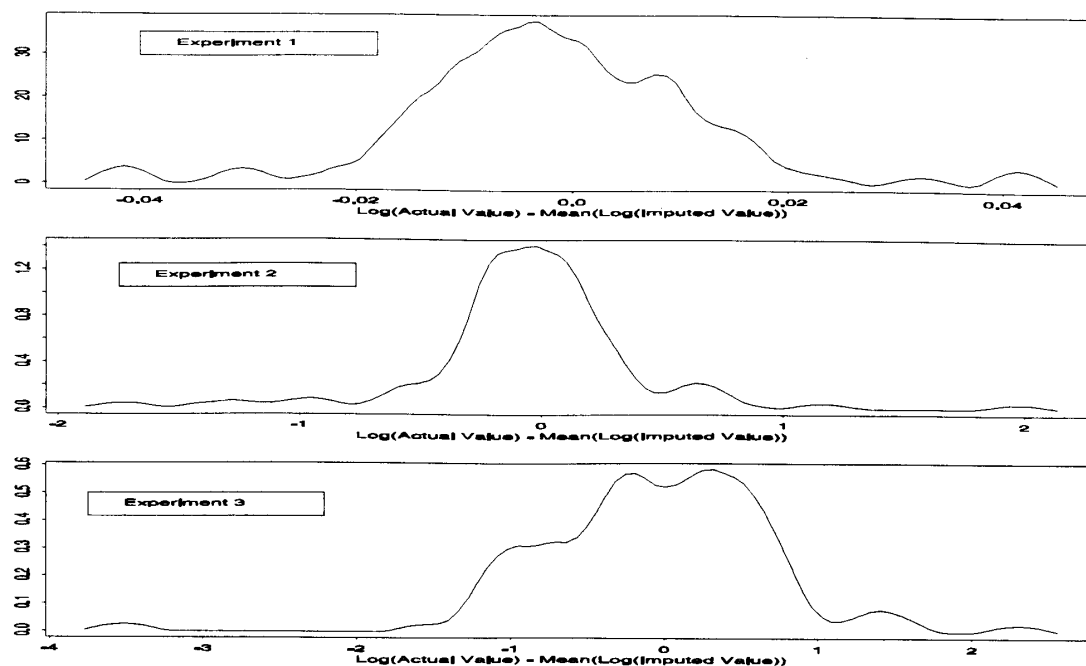


Figure 5b: ASH Plots of Distribution over Observations of the Standard Deviation of $\text{Log}_{10}(\text{Imputation})$ within Observations, Face Value of T-Bills, Experiments 1-3.

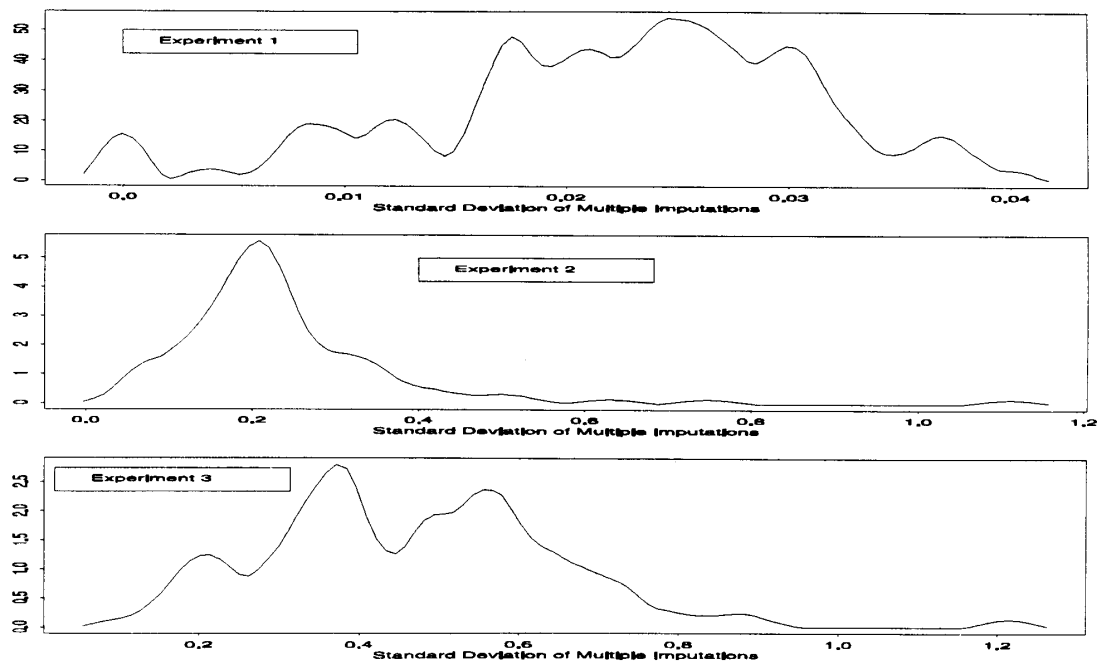


Figure 6a: ASH Plots of Distribution over Observations of $\text{Log}_{10}(\text{Actual Value}) - \text{Mean}(\text{Log}_{10}(\text{Imputation}))$ within Observations, Total Financial Assets, Experiments 1-3.

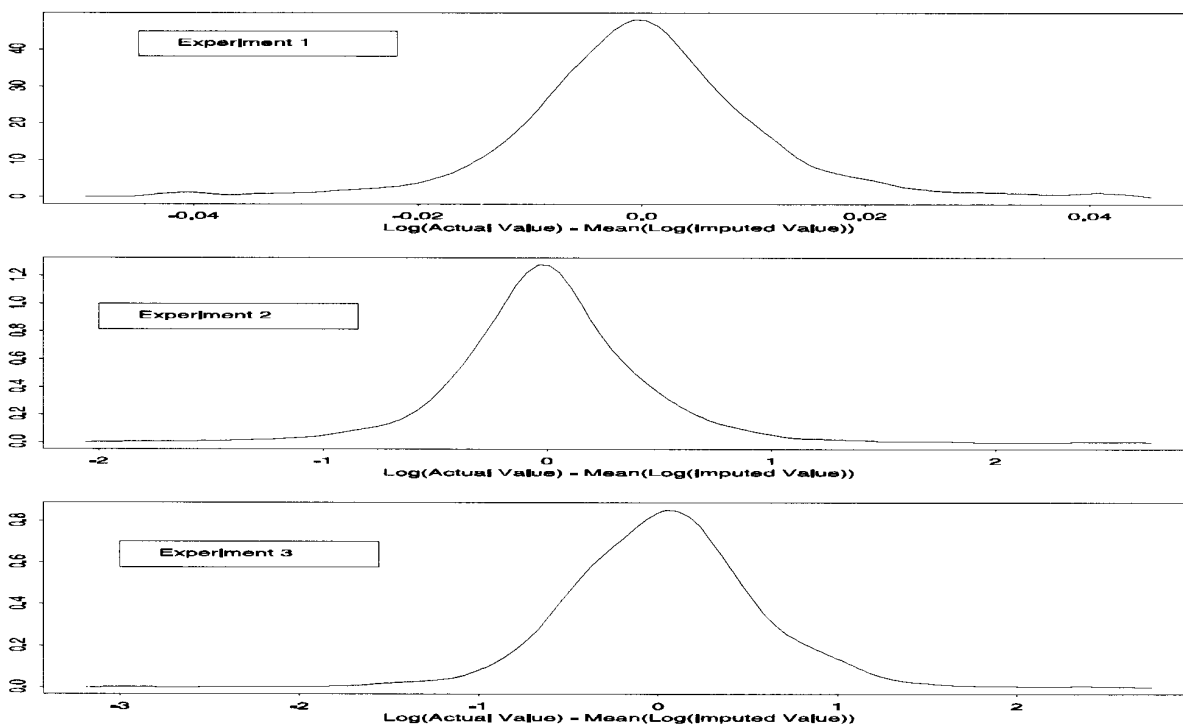
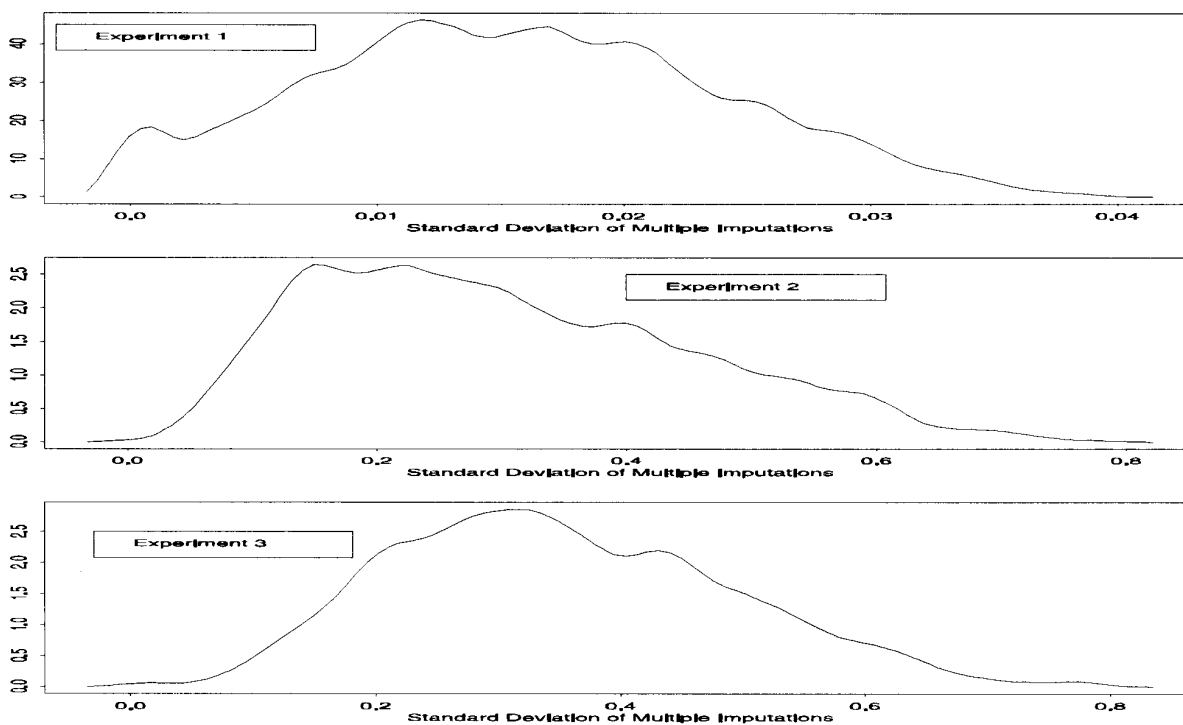


Figure 6b: ASH Plots of Distribution over Observations of the Standard Deviation of $\text{Log}_{10}(\text{Imputation})$ within Observations, Total Financial Assets, Experiments 1-3.



relatively “lumpy,” largely reflecting the smaller samples used to estimate these distributions: about 1,200 observations were used for the savings account estimate and only about 110 observations were used for the T-bill estimate, but about 2,900 were used to estimate the distribution for total income. Reflecting the integration over possibly many imputations, the distribution of bias for total financial assets is quite smooth. In every case shown, there is some piling up of cases at the outer bounds corresponding to ± 10 percent (about ± 0.04 on the log scale). The FRITZ model is allowed to draw as many as 400 times from the predicted conditional distribution of the missing data before selecting the nearest endpoint of the constraint. Thus, it is likely that these extreme observations are ones for which the models do not fit very well. Not surprisingly, examination of selected cases suggests that these observations are more likely to have unusual values for some of the conditioning variables in the imputation models. The median variability of the imputations within implicates shown by the ASH plots of the distributions of standard deviations, is about ± 6 percent for income, savings accounts, and T-bills. The variability within implicates is substantially lower for the sum of financial assets, reflecting offsetting errors in imputation.

In the second experiment, the relaxation of the simple range constraint in experiment one has the expected effect of increasing the variability of the bias, and increasing the standard deviation of imputations within implicates. In the case of total household income, the bias corresponding to the 90th percentile of the bias distribution jumps to about 25 percent. The effect is even larger for the other variables (the bias is nearly 300 percent at the 90th percentile for total financial assets). It is somewhat surprising just how much these values increase given that the imputations are potentially conditioned on a large number of reported values [15].

In the third experiment with the removal of the reported values used for conditioning in experiment two, the range of the bias rises further. The 90th percentile of the bias distribution is about 140 percent for total income, and about 400 percent for total financial assets.

Because these results are reported on a logarithmic scale, it is possible that they could be unduly influenced by changes that are small in dollar amounts, but large on a logarithmic scale. The data do not provide strong support for this proposition. For income, scatterplots reveal that the logarithmic bias appears to be approximately equally spread at all levels of income for experiments one and two [16]. In the third experiment, the dominant relationship is similar, but there are two smaller groups that deviate from the pattern: a few dozen observations with actual incomes of less than a few thousand dollars are substantially over-imputed on average, and a somewhat larger number of observations with actual incomes of more than \$100,000 are substantially under-imputed. The data suggest a similar relationships across the experiments for the other variables as well.

To gauge the effects of the experiments on the overall univariate distributions of the four variables considered, Figures 7-10 show quantile-quantile (Q-Q) plots of the mean imputations against the reported values on a logarithmic scale. Across these variables, the distribution is barely affected by experiment one. In the second experiment, the results are a bit more mixed. For total income and total financial assets, there is some over-imputation of values less than a few thousand dollars, and slight under-prediction at the very top. For T-bills, the relationship is much noisier, but not strikingly different. However, for savings accounts, the Q-Q plot is rotated clockwise, indicating that the imputed distribution is under-imputed at the top and over-imputed at the bottom. All of the simulated distributions deteriorate in the third experiment, though the distribution of total financial assets appears the most resilient [17].

Figure 7: Q-Q Plots of Imputed Distribution vs. Actual Distribution, Total Household Income, Experiments 1-3.

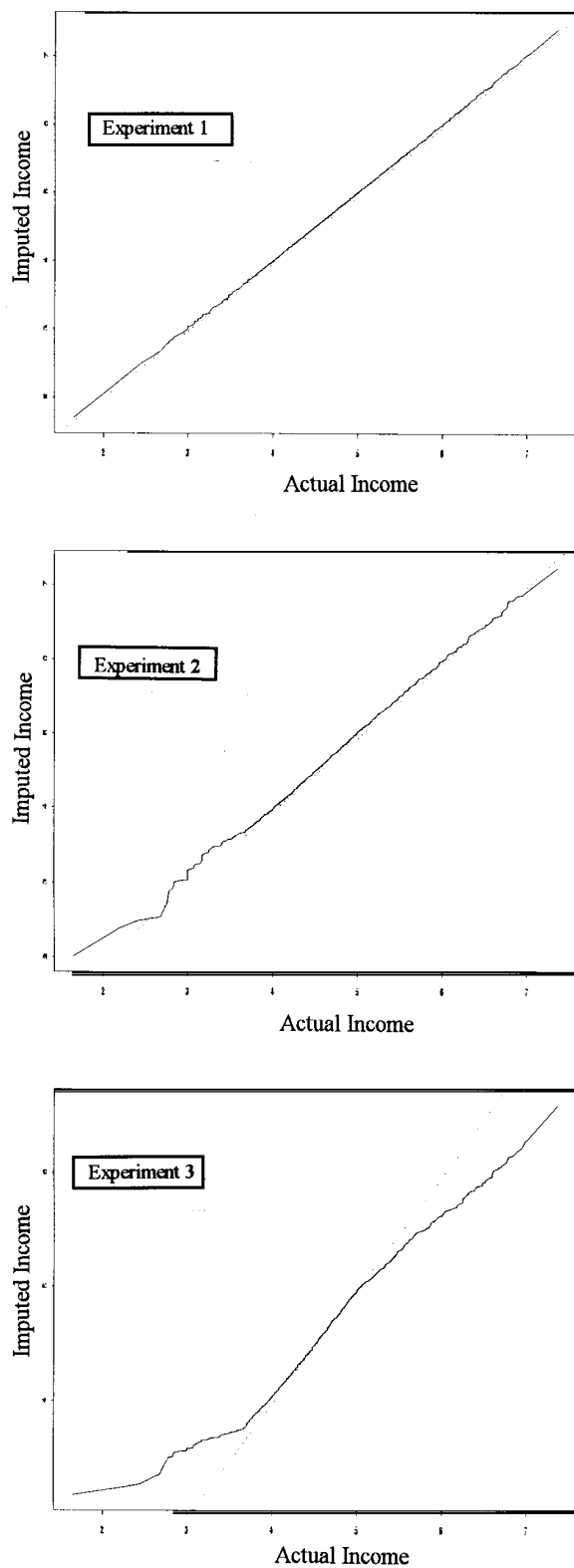


Figure 8: Q-Q Plots of Imputed Distribution vs. Actual Distribution, Balance in 1st Savings Account, Experiments 1-3.

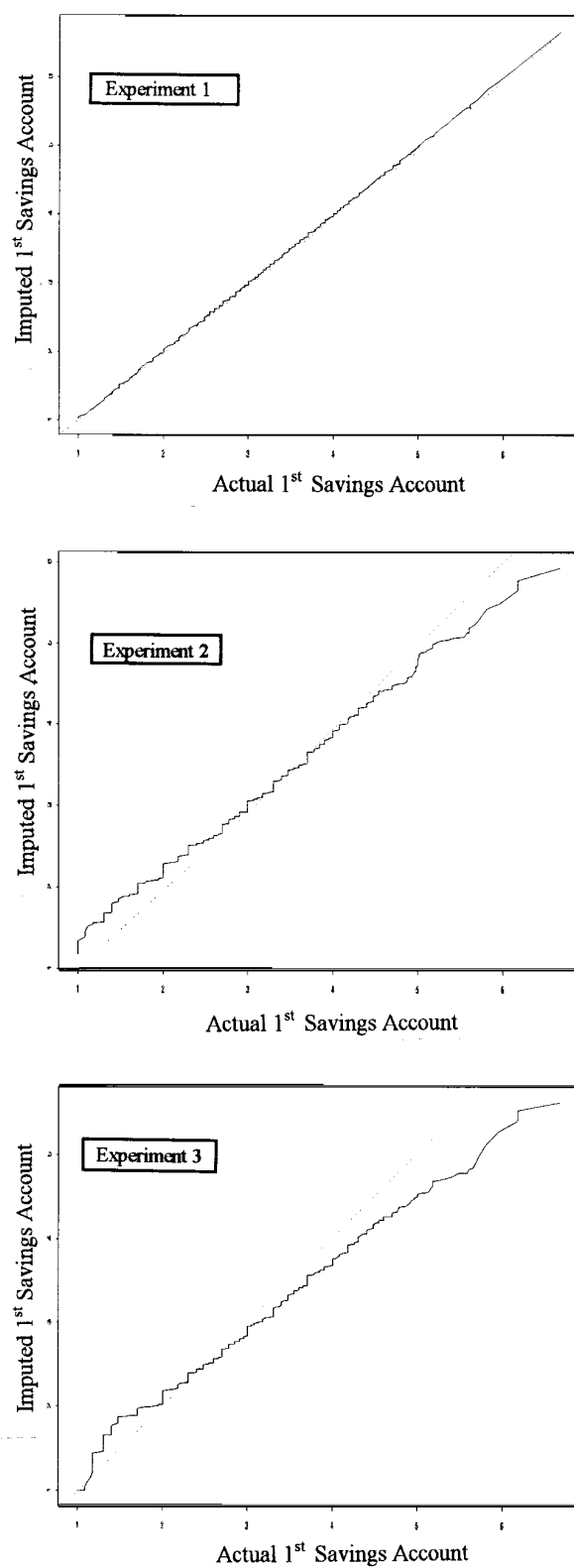


Figure 9: Q-Q Plots of Imputed Distribution vs. Actual Distribution, Face Value of T-Bills, Experiments 1-3.

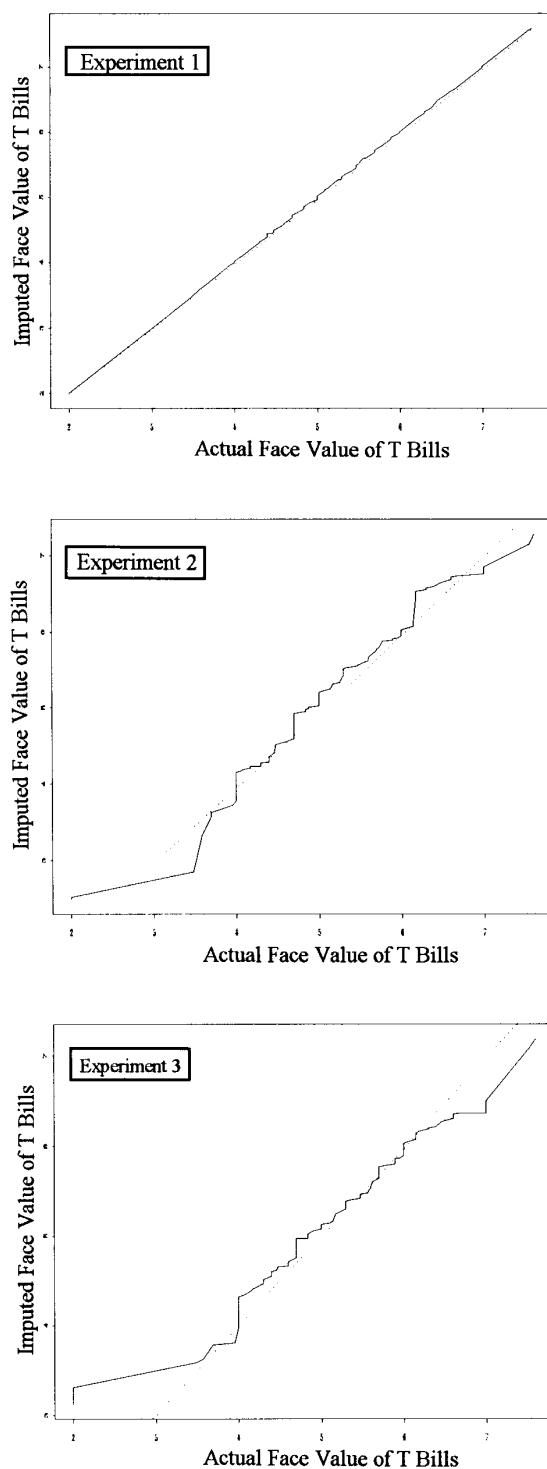
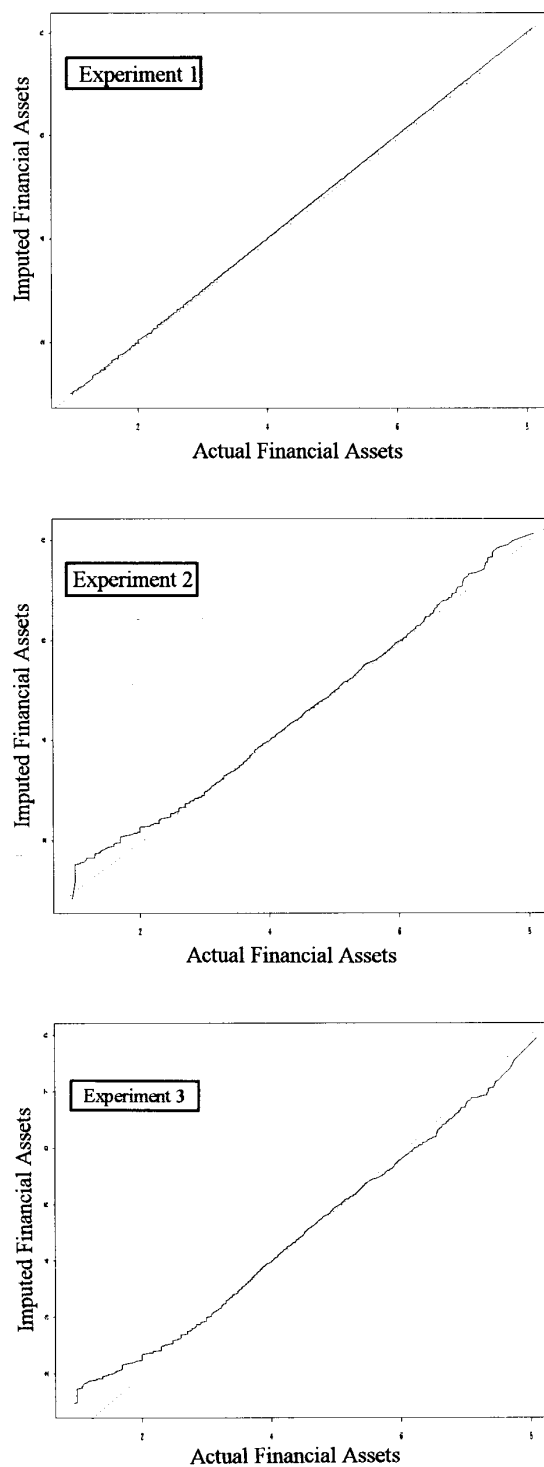


Figure 10: Q-Q Plots of Imputed Distribution vs. Actual Distribution, Total Financial Assets, Experiments 1-3.



Univariate and simple bivariate statistics are important for many illustrative purposes, but for the SCF, as is the case for many other surveys, the most important uses of the data over the long run are in modeling. Table 3 presents the coefficients of a set of simple linear regressions of the logarithm of total household income on dummy variables for ownership of various financial assets and the log of the maximum of one and the value of the corresponding asset. This model has no particular importance as an economic or behavioral characterization. It is intended purely as a descriptive device designed to examine the effects of the variation across the experiments on the partial correlations of a set of variables imputed in all the experiments. Two types of models are shown: one set includes all observations regardless of whether the variables included were originally reported completely by the respondent, and the other model includes only cases for which every variable in the model was originally reported completely. The regressions were run using data from each of the three experiments, as well as data from the final version of the sixth (final) iteration of the imputation of the main dataset [18].

Experiments one and two perform about equally well in terms of determining the significance of coefficients in both variations on the basic model. However, data from the first experiment misclassify one variable as not significant, and data from the second experiment misclassify some variables as significant. The third experiment implies both type one and type two errors. The R^2 of the regressions changes little except in the third experiment, where this value drops about 10 percent. Overall, none of the experiments do dramatically worse than the original data. Given the structure of the FRITZ model and the degree to which the variables in these regression models were mutually interdependent, it would be very surprising if the outcome were otherwise. However, such regressions are only the beginning of what many economist would consider applying to the data, and it is possible that more complex models or methods of estimation would give different results.

Summary and Future Research

By design, experiment one is virtually guaranteed to induce minimal distortions, but it also leaves the outcomes near the original values. Unfortunately, just knowing that an outcome is in a certain range may already be sufficient information to increase too much the probability of identifying some of the very wealthy respondents in the SCF. My *ex ante* choice of contenders among the experiments was the second one, in which imputations condition on actual values, but there is no prior constraint on the outcome that is connected to the original value. *Ex post*, I find the results relatively disappointing. Certainly, the reported outcomes of the third experiment look least attractive. There may be ways of more globally constraining or aligning the outcomes of experiments two and three, but I suspect the choice of method would depend critically on a ranking of the importance of the types of analyses to be performed with the data. I hope that someone in the SCF group or elsewhere will be able to take the next step.

One technical question that appears potentially troublesome is how to estimate sampling error in a fully simulated dataset [19]. It is possible, in theory, to simulate records for the entire universe, but even in this case there would still be sampling variability in the imputations. This variation may be a second order effect in normal imputation, but we need to deal with the issue carefully if we expect to simulate all the data. Perhaps we could find an approximate solution in independently multiply imputing each of a manageably small number of replicates — implicates of replicates; each replicate would require population estimates from a corresponding replicate selected from the actual data in a way that captured the important dimensions of variability in the sample. Another possibility might be to compute variance functions from the actual data.

Table 3. — Regression of Logarithm of Total Household Income on Various Variables, Original Data and Experiments 1-3, Using all Observations and Using Only Observations Originally Giving Complete Responses to all Variables in the Model

	All Observations Included				Only Complete Responders Included			
	Orig.	Exp. 1	Exp. 2	Exp. 3	Orig.	Exp. 1	Exp. 2	Exp. 3
Intercept	2.64*	1.92*	2.56*	3.76*	2.83*	2.87*	3.43*	6.60*
	<i>0.75</i>	<i>0.75</i>	<i>0.74</i>	<i>0.69</i>	<i>1.09</i>	<i>1.09</i>	<i>1.02</i>	<i>1.09</i>
Have checking	0.18*	0.20*	0.25*	0.21*	0.17*	0.18*	0.18*	0.15*
	<i>0.03</i>	<i>0.03</i>	<i>0.03</i>	<i>0.03</i>	<i>0.04</i>	<i>0.04</i>	<i>0.04</i>	<i>0.04</i>
Ln(\$ checking)	0.25*	0.27*	0.30*	0.26*	0.26*	0.27*	0.27*	0.23*
	<i>0.01</i>	<i>0.01</i>	<i>0.01</i>	<i>0.01</i>	<i>0.02</i>	<i>0.02</i>	<i>0.02</i>	<i>0.02</i>
Have IRA/Keogh	0.16*	0.18*	0.18*	0.17*	0.07	0.06	0.12	0.08
	<i>0.05</i>	<i>0.05</i>	<i>0.05</i>	<i>0.05</i>	<i>0.07</i>	<i>0.07</i>	<i>0.07</i>	<i>0.07</i>
Ln(\$ IRA/Keogh)	0.10*	0.11*	0.11*	0.10*	0.07*	0.07*	0.10*	0.08*
	<i>0.02</i>	<i>0.02</i>	<i>0.02</i>	<i>0.02</i>	<i>0.03</i>	<i>0.03</i>	<i>0.03</i>	<i>0.03</i>
Have savings acct.	0.01	0.02	0.01	0.01	-0.03	-0.03	-0.02	-0.03
	<i>0.04</i>	<i>0.04</i>	<i>0.04</i>	<i>0.04</i>	<i>0.04</i>	<i>0.05</i>	<i>0.05</i>	<i>0.05</i>
Ln(\$ savings acct)	0.03	0.03	0.03	0.04	0.00	0.00	0.01	0.01
	<i>0.02</i>	<i>0.02</i>	<i>0.02</i>	<i>0.02</i>	<i>0.02</i>	<i>0.02</i>	<i>0.02</i>	<i>0.02</i>
Have money market acct.	0.02	0.03	-0.04	-0.11	0.11	0.12	0.01	-0.07
	<i>0.07</i>	<i>0.07</i>	<i>0.07</i>	<i>0.07</i>	<i>0.09</i>	<i>0.10</i>	<i>0.10</i>	<i>0.10</i>
Ln(\$ money market acct.)	0.03	0.03	0.00	-0.02	0.05	0.05	0.01	-0.02
	<i>0.03</i>	<i>0.03</i>	<i>0.03</i>	<i>0.03</i>	<i>0.04</i>	<i>0.04</i>	<i>0.04</i>	<i>0.04</i>
Have CDS	0.24*	0.26*	0.31*	0.27*	0.22*	0.22*	0.27*	0.23*
	<i>0.08</i>	<i>0.08</i>	<i>0.08</i>	<i>0.08</i>	<i>0.10</i>	<i>0.11</i>	<i>0.11</i>	<i>0.11</i>
Ln(\$ CDS)	0.07*	0.07*	0.09*	0.08*	0.07	0.07	0.09*	0.07
	<i>0.03</i>	<i>0.03</i>	<i>0.03</i>	<i>0.03</i>	<i>0.04</i>	<i>0.04</i>	<i>0.04</i>	<i>0.04</i>
Have savings bonds	-0.02	-0.01	-0.05	-0.09	-0.10	-0.10	-0.12	-0.13
	<i>0.04</i>	<i>0.04</i>	<i>0.05</i>	<i>0.04</i>	<i>0.06</i>	<i>0.06</i>	<i>0.06</i>	<i>0.05</i>
Ln(\$ savings bonds)	0.02	0.02	0.00	-0.02	-0.03	-0.03	-0.05	-0.04
	<i>0.02</i>	<i>0.02</i>	<i>0.02</i>	<i>0.02</i>	<i>0.03</i>	<i>0.03</i>	<i>0.03</i>	<i>0.03</i>
Have other bonds	0.62*	0.65*	0.51*	0.63*	0.68*	0.66*	0.54*	0.35*
	<i>0.09</i>	<i>0.09</i>	<i>0.08</i>	<i>0.09</i>	<i>0.14</i>	<i>0.14</i>	<i>0.13</i>	<i>0.14</i>
Ln(\$ other bonds)	0.26*	0.27*	0.22*	0.25*	0.27*	0.26*	0.22*	0.15*
	<i>0.03</i>	<i>0.03</i>	<i>0.03</i>	<i>0.03</i>	<i>0.05</i>	<i>0.05</i>	<i>0.04</i>	<i>0.05</i>
Have mutual funds	0.06	0.07	0.09	-0.02	0.18	0.17	0.20*	0.00
	<i>0.07</i>	<i>0.07</i>	<i>0.07</i>	<i>0.05</i>	<i>0.09</i>	<i>0.09</i>	<i>0.09</i>	<i>0.06</i>
Ln(\$ mutual funds)	0.04	0.05	0.05*	0.01	0.10*	0.09*	0.10*	0.03
	<i>0.02</i>	<i>0.02</i>	<i>0.02</i>	<i>0.02</i>	<i>0.03</i>	<i>0.03</i>	<i>0.03</i>	<i>0.03</i>
Have annuity/trust	0.02	0.02	0.03	0.01	-0.04	-0.04	-0.07	-0.07
	<i>0.04</i>	<i>0.04</i>	<i>0.04</i>	<i>0.02</i>	<i>0.05</i>	<i>0.05</i>	<i>0.05</i>	<i>0.05</i>
Ln(\$ annuity/trust)	0.04*	0.04*	0.04*	0.02	0.01	0.01	0.01	-0.29
	<i>0.01</i>	<i>0.01</i>	<i>0.01</i>	<i>0.02</i>	<i>0.02</i>	<i>0.02</i>	<i>0.02</i>	<i>0.24</i>
Have whole life insurance	-0.70	0.11	0.14*	0.19*	-0.61	-0.63	0.17*	0.2*
	<i>0.17</i>	<i>0.08</i>	<i>0.05</i>	<i>0.05</i>	<i>0.25</i>	<i>0.26</i>	<i>0.07</i>	<i>0.06</i>
Ln(\$ cash value life ins.)	0.10*	0.01	0.02*	0.01	0.09*	0.09*	0.03	0.02
	<i>0.02</i>	<i>0.01</i>	<i>0.01</i>	<i>0.01</i>	<i>0.03</i>	<i>0.03</i>	<i>0.04</i>	<i>0.03</i>
R ²	0.40	0.39	0.40	0.37	0.43	0.43	0.42	0.36

* = significant at the 95% level of confidence.
Simple regression standard errors are given in italics below each estimate.

The experimental results reported in this paper say at least as much about the nature of the SCF imputations as they do about the possibility of creating a fully simulated dataset. Although the imputation models have been refined over three surveys now, the results of experiments two and three, in particular, suggest that there is room for improvement. Indeed, a number of changes were instituted in the process of getting the experiments to produce meaningful data, and other changes will be implemented during the course of processing the 1998 SCF. Other changes, including the possibility of using empirical residuals, deserve further attention. However, I am not optimistic that there are many major improvements in our ability to impute the SCF data waiting to be discovered. There is a difference in what one can accept in imputing a relatively small fraction of the data and what is acceptable for the whole dataset. With fully simulated data, we are left with a difficult tradeoff between noise (however structured) and potential disclosure.

Disclosure limitation techniques have a Siamese twin in record linkage techniques. As one side progresses, the other side uses closely related ideas to follow. This conference has played an important part in highlighting this relationship and the need for coordination. Perhaps if we work hard together, there may be a chance that we will find a way to allow users to analyze disclosure-limited data using record linkage ideas to sharpen inferences. There may also be a payoff in more routine statistical matching, which is really just another form of imputation.

A large problem in planning all disclosure reviews is how to accommodate the needs (but not necessarily all the desires) of data users. I expect that users will express considerable resistance to the idea of completely simulated data. Some statisticians may be troubled about how to address questions of estimating sampling error with such data. Among economists, there are substantial pockets of opposition to all types of imputation, and some researchers have raised carefully framed questions that need to be addressed equally carefully. For example, if unobserved effects are a serious issue (and they often are in econometric modeling), then imputation must consider the distortions it may induce if such latent models are ignored; the question becomes much more pressing if all of the data are imputed. Given the choice between having no data or having data that are limited in some way, most analysts will likely opt for some information. However, to avoid developing disclosure strategies that yield data that do not inform interesting questions for users, it may be important to engage users in the process where possible.

Acknowledgments

The author wishes to thank Kevin Moore and Amy Stubbendick for a very high level of research assistance. The author is also grateful to Gerhard Fries, Barry Johnson, and R. Louise Woodburn for comments, and to Fritz Scheuren for encouragement in this project.

Footnotes

- [1] As Fienberg (1997) argues, releasing any information discloses something about the respondent, even if the probability of identification is minuscule.
- [2] See Fries, Johnson and Woodburn (1997a) for a summary of the disclosure strategies that have been developed for the survey.
- [3] Ivan Fellegi emphasized a similar point in his address to this conferences.

- [4] For example, Rubin (1993) says “Under my proposal, no actual unit’s confidential data would ever be released. Rather, all actual data would be used to create a multiply-imputed synthetic microdata set of artificial units...”
- [5] However, Fienberg and Makov (1997) have proposed creating simulated data for the purpose of evaluating the degree of disclosure risk in a given dataset and Feinberg, Steele and Makov (1996) have examined the problem of simulating categorical data.
- [6] Use of the ITF for the SCF is strictly controlled to protect the privacy of taxpayers. For the 1995 SCF, SOI provided NORC with the names and addresses of a sample selected from a copy of the ITF purged of name and address information at the Federal Reserve. NORC contacted respondents, but had no means of linking to the tax data. The SCF group alone at the Federal Reserve is allowed access to both survey data and tax data, but no names were available, and use of these tax data at the Federal Reserve is strictly limited to activities connected with sampling, weighting, and other such technical issues.
- [7] See Fries, Johnson, and Woodburn (1997b) for details and information about the effects of the alterations on the data.
- [8] The shadow variables are used as a formal device in documentation, and they inform the imputation software about which variables should be imputed. The shadow variables contain information about various types of editing that may have been performed to reach the final value, whether it was reported as one of a large number of types of range outcomes, whether it was missing for various reasons, or whether its outcome was affected by other processes.
- [9] The collection of range data in the 1995 SCF is described in detail in Kennickell (1997).
- [10] For an excellent example of a simultaneously determined system, see Schafer (1995). Geman and Geman (1984) discuss another type of structure involving data “cliques.”
- [11] In general, continuous variables are assumed to follow a conditional lognormal distribution. For continuous variables, the program assumes by default that errors should be drawn within a bound of 1.96 standard errors above and below the conditional mean.
- [12] For the 1995 data, the process required about ten days per iteration, which is down from about four weeks per iteration in 1989.
- [13] There are 480 monetary variables in the SCF, but it is not possible for a given respondent to be asked all of the underlying questions.
- [14] The sets of observations underlying the charts include only respondents who gave a complete response for the variable, or, in the case of financial assets, who gave complete responses for all the components of financial assets. For many sub-models of the SCF implementation of FRITZ, general constraints are imposed for all imputations to ensure values that are reasonable (e.g., amounts owed on mortgage balloon payments must be less than or equal to the current amount owed); in the actual data, these constraints are occasionally violated for reasons that are unusual, but possible. When reimputing these values subject to dollar range constraints in experiment one, a small number of imputations violated the bounds imposed. To avoid major restructuring of the implementation of the FRITZ model for the experiments, these instances are excluded from the comparisons reported here. In each of the figures, the set of observations is the same across all six of the panels. For the income plots, households reporting negative income have been excluded.

- [15] For example, total income is the first variable imputed, and all reported values (or midpoints of ranges) for variables included in the model for that variable are used to condition the imputation.
- [16] For disclosure reasons, the scatterplots supporting this claim cannot be released.
- [17] In the cases examined, this result also holds if the data are separated by implicates rather than averaged across implicates.
- [18] The five implicates were pooled for these regressions. Standard errors shown in the table are simple regression standard errors that take no account of imputation or sampling error; the degrees of freedom were altered in the standard error calculation to reflect the fact that there were five times as many implicates as observations.
- [19] Fienberg, Steele and Makov (1996) also address this question.

References

- Fienberg, Stephen E. (1997). Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistics Research, working paper, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.
- Fienberg, Stephen E. and Makov, Udi E. (1997). Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data, working paper, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.
- Fienberg, Stephen E.; Steele, Russell J.; and Makov, Udi E. (1996). Statistical Notions of Data Disclosure Avoidance and their Relationship to Traditional Statistical Methodology: Data Swapping and Loglinear Models, *Proceedings of the 1996 Annual Research Conference and Technology Interchange*, Washington, DC: U.S. Bureau of the Census, 87-105.
- Fries, Gerhard; Johnson, Barry W.; and Woodburn, R. Louise (1997a). Analyzing Disclosure Review Procedures for the Survey of Consumer Finances, paper for presentation at the 1997 Joint Statistical Meetings, Anaheim, CA.
- Fries, Gerhard; Johnson, Barry W.; and Woodburn, R. Louise (1997b). Disclosure Review and Its Implications for the 1992 Survey of Finances, *Proceedings of the Section on Survey Research Methods*, 1996 Joint Statistical Meetings, Chicago, IL.
- Geman, Stuart and Geman, Donald (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6, 6 (November), 721-741.
- Kennickell, Arthur B. (1991). Imputation of the 1989 Survey of Consumer Finances, *Proceedings of the Section on Survey Research Methods*, 1990 Joint Statistical Meetings, Atlanta, GA.
- Kennickell, Arthur B. and Woodburn, R. Louise (1997). Consistent Weight Design for the 1989, 1992 and 1995 SCFs, and the Distribution of Wealth, working paper, Board of Governors of the Federal Reserve System, Washington, DC.
- Kennickell, Arthur B. (1997). Using Range Techniques with CAPI in the 1995 Survey of Consumer Finances *Proceedings of the Section on Survey Research Methods*, 1996 Joint Statistical Meetings, Chicago, IL.

- Little, Roderick J.A. (1983). The Nonignorable Case, *Incomplete Data in Sample Surveys*, New York: Academic Press.
- Rubin, Donald B. (1993). Discussion of Statistical Disclosure Limitation, *Journal of Official Statistics*, 9, 2, 461-468.
- Schafer, Joseph (1995). *Analysis of Incomplete Multivariate Data*, Chapman and Hall.
- Tourangeau, Roger; Johnson, Robert A.; Qian, Jiahe; Shin, Hee-Choon; and Frankel, Martin R. (1993). Selection of NORC's 1990 National Sample, working paper, National Opinion Research Center at the University of Chicago, Chicago, IL.

The views presented in this paper are those of the author alone and do not necessarily reflect those of the Board of Governors or the Federal Reserve System. Any errors are the responsibility of the author alone.